

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Nástroj pro automatickou identifikaci KIR alel

Místo této strany bude
zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 19. března 2020

Kateřina Kratochvílová

Poděkování

Ráda bych poděkovala Ing. Lucii Houdové, Ph.D. za cenné rady, věcné připomínky, trpělivost a ochotu, kterou mi v průběhu zpracování této práce věnovala.

Abstract

The text of the abstract (in English). It contains the English translation of the thesis title and a short description of the thesis.

Abstrakt

Text abstraktu (česky). Obsahuje krátkou anotaci (cca 10 řádek) v češtině. Budete ji potřebovat i při vyplňování údajů o bakalářské práci ve STAGu. Český i anglický abstrakt by měly být na stejné stránce a měly by si obsahem co možná nejvíce odpovídat (samozřejmě není možný doslovný překlad!).

Obsah

1 Úvod	8
2 Geny	9
2.1 Imunitní systém aneb kde se vzal KIR	10
2.1.1 Vrozená imunita	10
2.1.2 Získaná imunita	10
2.1.3 Antigen	10
2.2 HLA a non-HLA geny	10
2.2.1 KIR	13
2.2.2 MICA/MICB	13
2.3 Porovnání vhodného dárce	14
2.4 Natural Killer a KIR	15
2.4.1 Natural killer	16
2.4.2 KIR	16
2.4.3 Nomenklatura KIR genů	19
2.5 KIR haplotyp	21
2.5.1 Dědičnost KIR	23
2.6 Sekvence DNA	23
2.7 Alela a gen	24
2.7.1 Alely KIR genů	25
2.8 Bordel - KIR	25
2.9 bordel	26
2.9.1 bordel hla	26
2.9.2 bordel zbytek	27
2.10 Bordel pro první kapitulu	29
3 Sekvenační metody získávání DNA dat	31
3.1 Sanger sequencing	31
3.2 NGS next-generation sekvenování	32
3.2.1 454 sekvenování a Ion Torrent	33
3.2.2 Illumina	33
3.2.3 SOLiD	34
3.3 Metody třetí generace	34
3.4 Read	36
3.5 Bordel	37

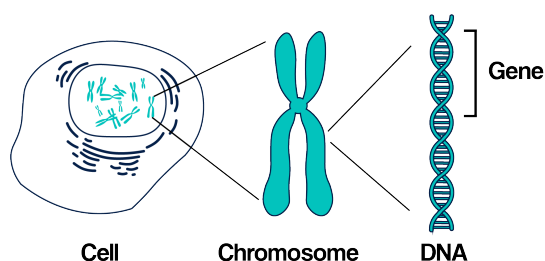
4	Analyza dostupných bioinformatických nástrojů pro zpracování NGS dat	40
4.1	Referenční geny	40
4.1.1	Vytvoření testovacího haplotypu	40
4.2	ART	41
4.2.1	pokus to nějak spustit	41
4.2.2	FASTQ	42
4.2.3	bordel	43
4.3	Bowtie	44
4.3.1	Bordel	44
4.3.2	Bowtie 2	45
4.3.3	bordel	45
5	Seznam zkratk	47
6	Výkladový slovník pojmů?	48
	Literatura	49

1 Úvod

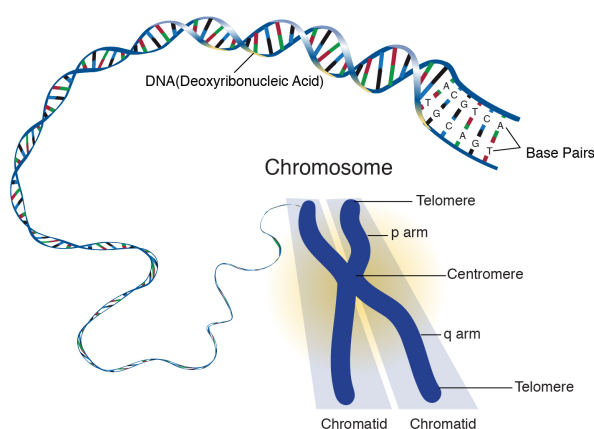
něco o transplantaci kostní dřeně, narazit že když by bylo 10/10 tak by bylo vhodné dělat porovnání těch KIR něco o AML Obtížnost transplantace krvetvorných buněk nespočívá v lékařské části procesu, nýbrž v následné imunitní odpovědi organismu. Proč KIR? Protože roste počet důkazů vlivu genů KIR že mají vliv na výsledky transplantace při leukemii

2 Geny

V každé buňce lidského organismu, konkrétně v buněčném jádře, je možné nálezt 46 chromozomů. Jeden chromozom představuje stočenou dlouhou molekulu DNA (Deoxyribonuklenovou kyselinu). Všechny 46 chromozomů obsahuje okolo 100 000 genů. Drobný segment DNA, který řídí buněčnou funkci je právě gen. Konkrétní forma genu je alela. [19]



Obrázek 2.1: Gen a vztah k lidské buňce. [4]



Obrázek 2.2: Struktura chromozomu. Na tomto obrázku je důležité si povšimnout p a q raménka a centromeru. [1]

Uvnitř buňky máme celý genom který se ovšem nemusí projevit na povrchu buňky. Pokud se vlastnost kterou gen přenáší projeví na povrchu buňky označujeme to jako exprese genu (jeho sebevyjádření). Od toho se odvíjí i označování něco za KIR gen či KIR receptor nebo molekulu.

2.1 Imunitní systém aneb kde se vzal KIR

Imunitní systém chrání organismus před škodlivinami. Skládá se ze dvou hlavních částí vrozené imunity a získané imunity.

2.1.1 Vrozená imunita

Vrozená imunita též označována přirozená, neadaptivní, antigenně nespecifická je neměnně zapsána v DNA. To znamená, že při každém setkání s antigenem odpoví stejnou reakcí. Buňky nesoucí vrozenou imunitu jsou stále přítomně v krvi, takže jejich případná aktivace je takřka okamžitá (minuty až hodiny). Do této imunity patří i natural killer buňky s KIR receptory, které budou dále rozebírány v textu.

2.1.2 Získaná imunita

Získaná imunita též označována specifická či adaptivní oproti specifické má v genomu zapsány pouze své základy. V průběhu lidského života se vyvíjí a mění. Změna může nastat například očkování nebo proděláním příslušné choroby. Tato změna ovšem nemusí být trvalá. Z těchto důvodů může být odpověď získané imunity při setkání se stejnou chorobou rozdílná. Fungování získané imunity zajišťují T- a B- lymfocyty, ale nefunguje samostatně. Při zabíjení patogenů spolupracuje s vrozenou imunitou.

2.1.3 Antigen

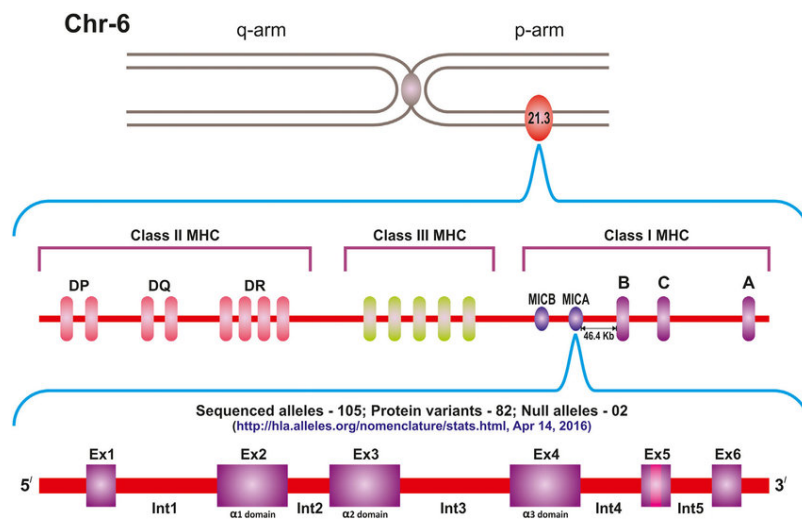
Antigeny jsou látky, které imunitní systém rozpozná a zareaguje na ně. V podstatě to může být jakákoli bílkovinná sloučenina. Antigen se obvykle nachází na povrchu buňky jako vyjádření genu. Imunitní systém následně zjistí o jaký antigen se jedná, respektive o jakou buňku se jedná, zda tělu vlastní (např. zdravá buňka) nebo buňku tělu cizí (např. nádorová buňka), tedy jedná-li se o exprese lidského genu nebo například viru. Jedná-li se o buňku tělu cizí imunitní systém reaguje snahou ji zničit.

2.2 HLA a non-HLA geny

Human leucocyte antigen (HLA) je genetický systém, který je primárně zodpovědný za rozeznávání vlastního od cizorodého. Někdy je termín HLA zaměňován s MHC. MHC (Major histocompatibility complex) je souhrnný termín pro všechny komplexy, kdy podskupinou jsou právě HLA (H - Human)

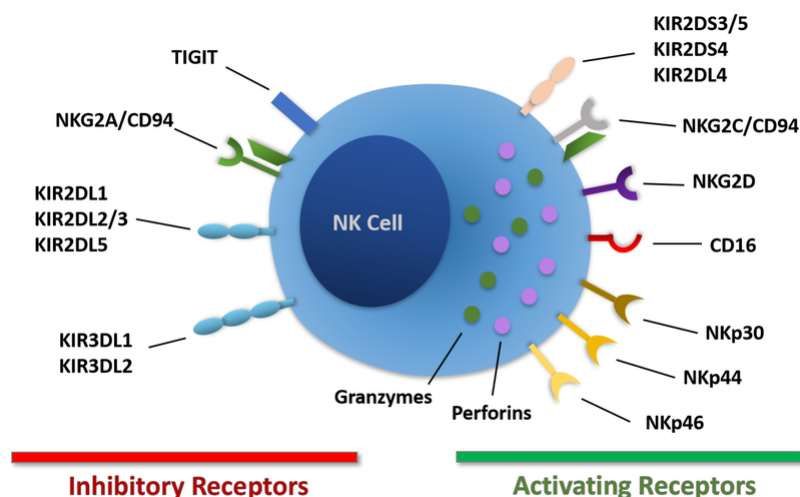
který je pro lidi. Stejně tak existuje DLA (D - Dog) který je pro psy. Z funkčního i biologického hlediska jde však u všech savců o stejnou skupinu genů

HLA a některé non-HLA geny se nacházejí na krátkém raménku 6 chromozomu, konkrétně 6p21.3 a zaujímá úsek přibližně jednu tisícinu genomu. Tento region je nejvíce komplexní a polymorfní na lidském genomu s více než 220 geny. Oproti tomu jedna ze skupin non-HLA genů, konkrétně KIR geny, se nachází na 19 chromozomu. Rozsáhlá diverzita genů vznikala snahou eliminovat neustále se měnící spektrum patogenů. Produkty těchto genů na povrch buňky významně ovlivňují odpověď na infekční choroby a výsledky buněčné či orgánové transplantace. [16]



Obrázek 2.3: 6 chromozom. TODO ten spodní řádek bych asi usekla- stejně nevím co to znamená.

Přesná definice mezi HLA a non-HLA geny neexistuje. Mimo jiné i jejich rozdělení není v literaturách sjednocené. Jak je vidět z obrázku 2.4 je možné geny rozdělit do tří tříd. V některých literaturách je možné nalést označení non-HLA genů jako geny III.třídy v jiné, že jsou to všechny geny III třídy a některé geny třídy I. Tato práce se bude v označení za gen non-HLA či HLA odkazovat na hla.alleles [16]. Zjednodušeně tedy můžeme říci, že geny které nejsou řazeny k HLA skupinám jsou non-HLA. Je-li gen označen za non-HLA neznamena to, že by neměl souvislost s funkcí imunitního systému. Naopak má, jen ne výlučně s HLA systémem. Non-HLA geny kódují produkty spojené s imunitními procesy. Mezi non-HLA geny mimo jiné patří MICA, MICB a KIR. [16]



Obrázek 2.4: Natural killer buňka a její receptory, rozděleny na aktivační a inhibiční. Pro tuto práci jsou důležité hlavně KIR receptory a NKG2D. [5]

TODO2 - obecně bych zkrátila popis HLA a trochu rozšířila non-HLA - stačí KIR a pak MICA, MICB jako ligandy NKG2D, a to jenom základ, protože podrobně KIRy máte dál, stačí asi jen zpřeházet odstavce, využít zde nějakou základní info o KIR a NK buňkách, kterou máte až dál nebo v té části bordel :) jako KIR 19 chromozom, NKG2D nebude to natural killer?

TODO neměla bych někde popsat co je to ligand? Já upřímně sama moc nevím co si pod tím mám představit Ligandy (Ligand = částice (atom, molekula, iont), která se váže na centrální atom, řadí se abecedně Centrální atom = částice, na kterou se vážou ligandy Centrální atom poskytuje volné orbitály elektronovým párům ligandů. Ligand je atom, ion nebo molekula, poskytující jeden nebo více elektronových párů centrálnímu atomu. Ligand je součástí komplexních (koordinačních) zlúčenin. Každá komplexná zlúčenina obsahuje centrální atom (kation) a anion nebo neutrální komplex (ligand) -wikiskripta

Ligand ve smyslu používaném v biochemii a farmakologii označuje látku, typicky malou molekulu, která vytváří komplex s biomolekulou a tato vazba má biologický význam. V užším slova smyslu se jako ligand označuje signální molekula, která se váže na vazebné místo cílového proteinu. Ligand, který je schopný po navázání na receptor vyvolat fyziologickou odpověď, se nazývá agonista, ten, který je schopen se vázat, ale odpověď nespouští, je antagonist

- wikipedie

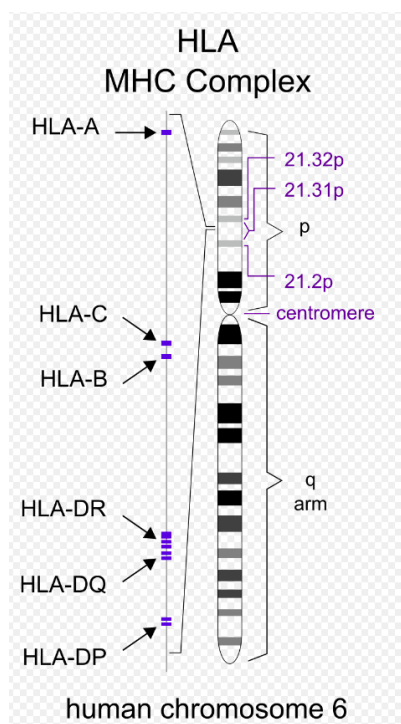
2.2.1 KIR

Killer immunoglobulin-like receptor (KIR) je skupina genů řazených mezi non-HLA geny. Jejich expresí jsou receptory na povrchu natural killer buněk (NK). Dnes je známo 15 genů rozlišujících se na inhibiční a aktivační na základě cytoplasmatického ocásku a počtu imunoglobulinových domén. TODO Možná dopsat že se váží na HLA-I ligandy TODO možná zmínit že jsou tam ještě ty pseudogeny

2.2.2 MICA/MICB

NKG2D se zase váže na MICA/MICB NKG2D je receptor na NK buňce a ještě je tam nějaký natural cytotoxic receptor (NCRs) receptor rozpoznávající stress-damage cells můžou spustit cytotoxicitu i když jsou tam inhibiční a HLA-I ligandy Ligandy pro NKG2 jsou MICA, MICB, ULBPs TODO možná zmínit co je NKG2

Geny skupiny MICA a MICB jsou dle hla.alleles [16] označeny jako class I chain-related gene. To znamená, že se běžně neřadí do I. třídy, i když je možné toto označení v některých literaturách nalézt. Oproti HLA genům, které mají svoje produkty na lymfocytech, se produkty MICA a MICB nachází na epitelových buňkách. Nejedná se tedy o standardní HLA geny, proto jsou nověji v literaturách označovány jako non-HLA. Jejich expresí na povrch buňky jsou ligandy, které se váží na receptor NKG2D. Buňky s ligandy MICA a MICB se množí při nádorovém onemocnění, zanětu nebo pod vlivem různých forem buněčného stresu a díky navázáním na receptor může být spuštěna imunitní reakce. [15] [7]



Obrázek 2.5: TODO ještě jsem sem chtěla dát obrázek, jenže je wiki original ale přijde mi hodně hezký Gen a vztah k lidské buňce. I když ten nad ním se mi možná líbí víc když je tam MICA MICB [4]

2.3 Porovnání vhodného dárce

TODO tak nevím jestli to teda nedat až za KIR?

Aby mohl být člověk dárce musí být ve věku 18-35 let a mít dobrý zdravotní stav. Při určování shody dárce a pacienta se rozhoduje na základě shody alel u genů HLA -A, -B, -C, -DRB1, -DQB1. Díky velké diverzitě HLA genů je počet možných kombinací několik miliard. Některé kombinace genů se vyskytují na základě oblasti či národnosti častěji nebo mohou být naopak vzácné. HLA geny se obvykle dědí jako blok (celý haplotyp), avšak ve výjimečných případech může dojít k rekombinaci. Z tohoto důvodu je nejsnadnější nalézt shodu v pokrevním příbuzenstvu.

Jelikož každý jedinec má dvakrát geny na pozicích HLA -A, -B, -C, -DRB1 a -DQB1 (jednu pěticí od otce, druhou pěticí od matky), je maximální shoda 10/10 (shoda obou alel v lokusech). Čím je shoda menší tím větší je riziko nepřijetí stěpu. U nepříbuzných jedinců lze tolerovat shodu 9/10 či 8/10. [6]

[13]

V posledních letech se objevuje Haploidentická transplantace, kdy je možné použít krvetvorné buňky příbuzného pouze s poloviční shodou (např. všichni rodiče a děti). Umožňuje to podávání chemoterapie pár dní po transplantaci, která zničí všechny buňky, které tělo nepřijme. Využívá se toho hlavně v případech časové tísně, kdy není čas hledat dárce v registrech. [3]

HLA-6.Chromozom a KIR 19.chromozom udíží se segregují nezávisle a HLA shodní dárce s příjemce mají obvykle různé složení KIR genů (Fryčová) TODO2 + dodat informaci o sekundárních kritériích, kde kromě věku apod. se řeší právě KIR haplotyp a v návaznosti se výzkumně zjišťuje, zda může mít vliv nejen KIR haplotyp - takže i dědičnost KIR uvést, ale i KIR alely samotné

TODO2 možná uvést někde rozdíl genotyp, fenotyp

TODO2 U nepříbuzných jedinců lze tolerovat i shodu 9/10 či 8/10. V případě vyšší shody u nepříbuzenského dárce oproti shody u příbuzného dárce může být upřednostněn příbuzný. dnes už trochu jinak - viz haploidentické transplantace (najděte si info na webu)

K typizaci se nejčastěji používá PCR – SSP (PCR se sekvenačně specifickými primery) či SBT (sequence based typing) technika, v posledních letech se stává zlatým standardem přímá sekvenace (SBT) HLA genu. není sepsáno příloha

2.4 Natural Killer a KIR

Mezi rizika při transplantaci krvetvorných buněk patří reakce štěpu proti hostiteli nebo relaps onemocnění (návrat nemoci). Podle nedávných studií výsledky přijetí štěpu ovlivňují nejenom HLA geny ale i non-HLA geny. Jedním z nich může být právě killer immunoglobulin-like receptor (KIR). V případech kdy by bylo nalezeno více vhodných dárců, tj. se shodou 10/10 nebo 9/10, vybíralo by se následně podle KIR genů [13] [6].

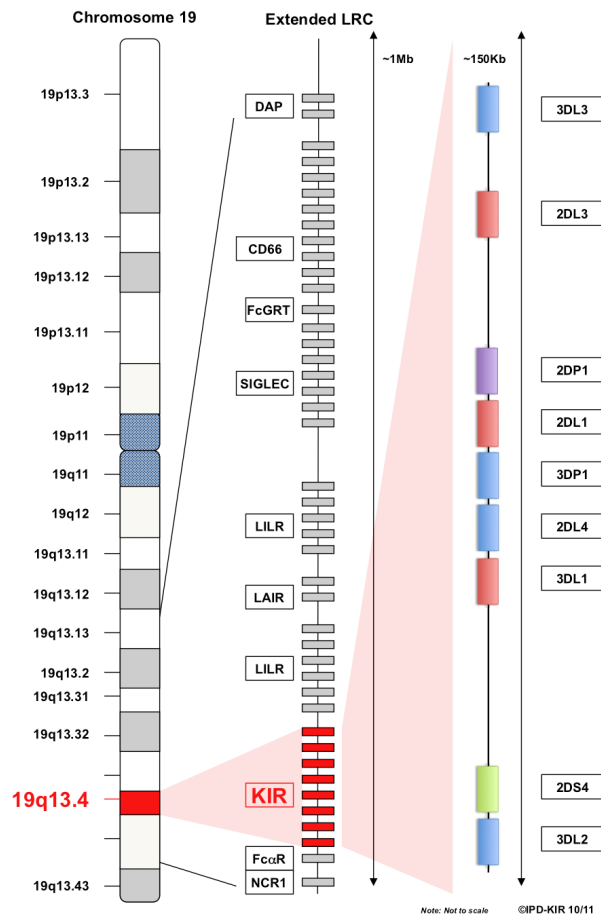
2.4.1 Natural killer

NK buňky (Natural killer) jsou velké granulární lymfocyty vrozeného imunitního systému. V krevním oběhu lidského těla je jich možné nalést 10–15%.

Klíčovou vlastností NK buněk je nejenom schopnost rozlišit poškozené buňky od zdravích ale i poškozené buňky rychle a efektivně likvidovat. Poškozené buňky mohou být buňky infokované virem či buňky transfomované v nádorové.

2.4.2 KIR

Killer immunoglobulin-like receptor (KIR) jsou receptory na povrchu NK buněk. Natural killer buňky oproti B- a T- lymfocitům (buňkám získané imunity) nemají antigenně specifické receptory. NK buňky rozpoznávají a zabíjejí buňky na základě interakce mezi KIR receptorem a HLA molekulou na povrchu zkoumané buňky.



Obrázek 2.6: KIR se nachází na 19 chromozomu. [16]

TODO2 určitě zmínit zde, že jich je 16, které jsou které S x L a jak jsou definovány KIR haplotypy A a B TODO 16 není jich 15? Nebo jinak nevím čeho 16

TODO 2 trochu popřeskládat a propojit s tím, co máte uvedeno dále v nomenklatuře

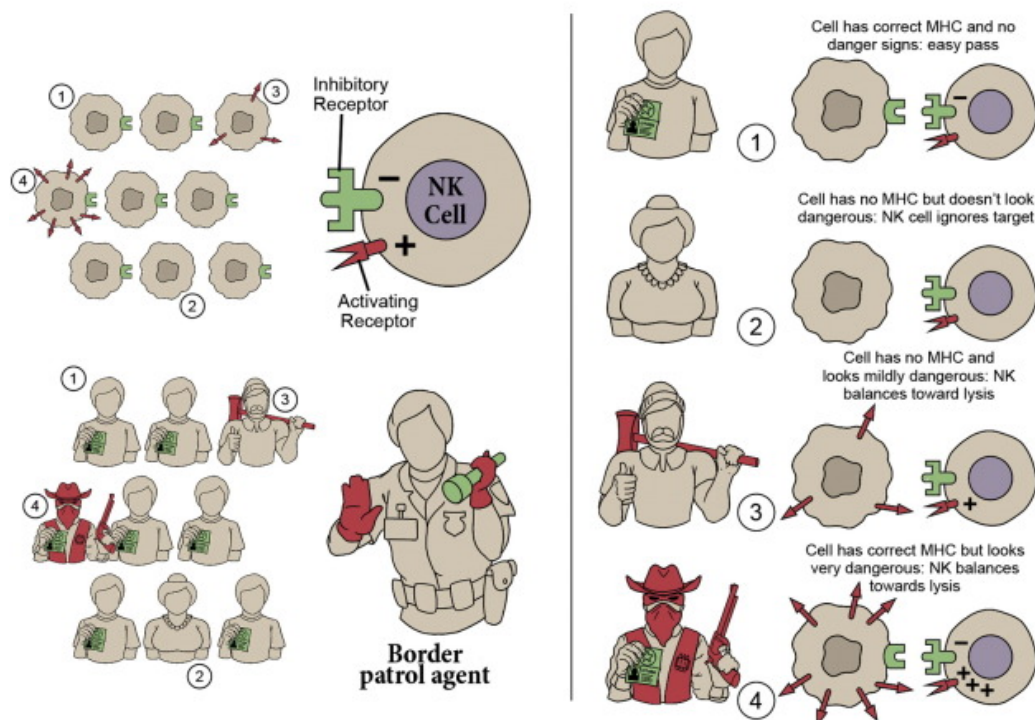
KIR receptory můžeme rozdělit na inhibiční a aktivační. Zda dojde k aktivaci NK buňky rozhoduje právě jejich rovnováha na zkoumané buňce. Zatímco inhibiční receptory se váží hlavně na molekuly HLA, aktivační receptory rozpoznávají molekuly, které jsou exprimovány na membránu při buněčném stresu.

TODO2 za tohle dát obrázek 2.7, takže možná tu nomenklaturu vložit někam mezi tyto první odstavce

NK buňky ustavičně prohledávají své okolí a testují přítomnost příslušných HLA ligand (specifická HLA molekula) pro své KIR receptory. Pokud je příslušný HLA ligand přítomen naváže se na NK buňku (2.7 případ 1). Tímto systémem jsou ochráněny vlastní HLA buňky. Pokud přítomen není je spuštěna cytotoxická reakce (schopnost ničit buňky) a zkoumaná buňka je zničena.

Některé virem napadené buňky potlačují propsání HLA ligand na povrch buňky a tím se brání cytotoxicitě proti T lymfocitům, ale naopak jsou více citlivější na cytotoxicitu proti NK buňkám. Ukázáno na obrázku 2.7 případ 3.

The NK Cell is like a border patrol agent



Obrázek 2.7: Přirovnání fungování natural killer buňky k pasové kontrole [18]. V pravé části jsou zobrazené případy které mohou nastat když natural killer buňka potká jinou buňku. V prvním případě je tělu vlastní zdravá buňka, kde se KIR receptor naváže na HLA ligand a k cytotoxické reakci nedojde. V druhém případě je červená krvinka k cytotoxicitě opět nedojde, protože na zkoumané buňce nepřevažují aktivační receptory. V 3 případě je to nádorová buňka, která schová HLA ligand (může nastat po transplantaci kostní dřeně) a tím se "schová" proti T- lymfocytům. Avšak aktivační receptory převládají a tak k cytotoxicitě dojde. Ve 4 příkladě je nádorová buňka nebo virem nakažená buňka (stresové ligandy). Aktivační receptory převládají k cytotoxicitě dojde.

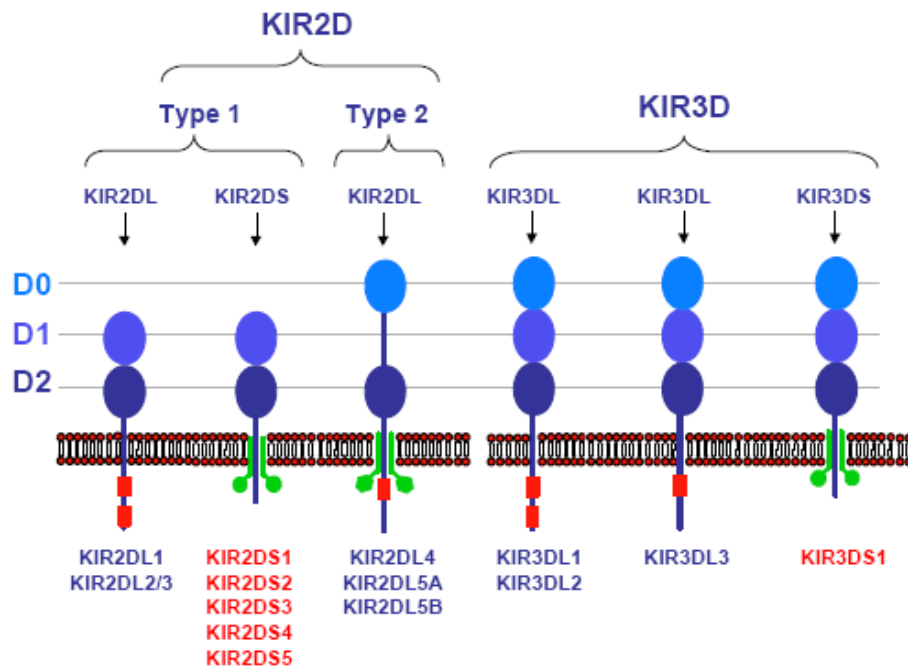
TODO tohle si nejsem jistá, že je pravda V případě transplantace je jedním z možných řešení pacientovi dát štěp, který obsahuje aktivační KIR receptory, které nemají jeho buňky.

2.4.3 Nomenklatura KIR genů

KIR geny (na obrázku 2.8 se liší různou délkou cytoplasmatických ocásku (tail) a různým počtem imunoglobulin-like domén (lg-like). Na základě této

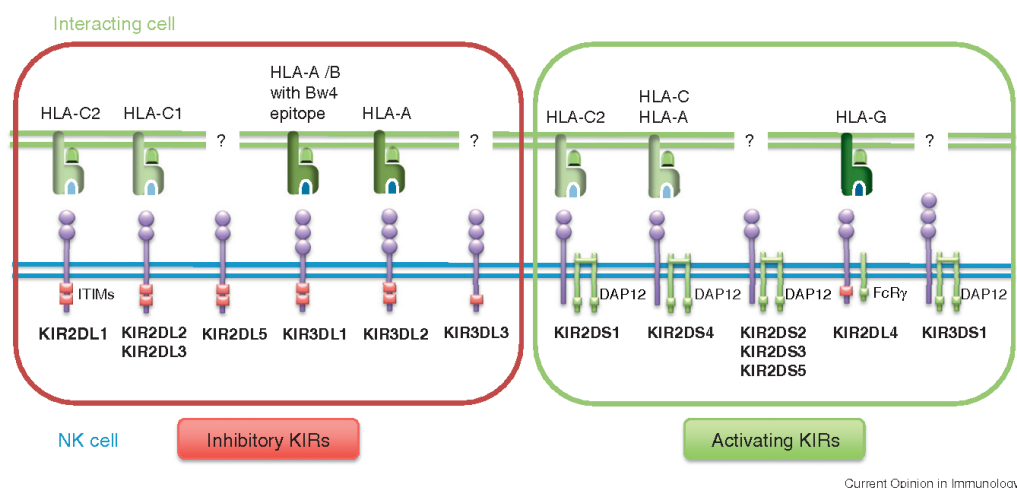
rozmanitosti byla založena nomenklatura KIR genů, tedy jejich pojmenování.

Jak je vidět na obrázku 2.8, cytoplasmatický ocásek může být dlouhý (long - L) nebo krátký (short - S). Oproti tomu imunoglobulinové domény se mohou vyskytovat 2 (2D) nebo 3 (3D).



Obrázek 2.8: Nomenklatura KIR genů. [13]

Další rozdělení KIR genů je již výše zmíněné inhibiční a aktivační. Na obrázku 2.9 je možné si povšimnout detailu, že až na KIR2DL4 jsou aktivační KIR s krátkým ocáskem, zatímco inhibiční jsou s dlouhým ocáskem. Obrázek dále uvádí vazebné ligandy pro jednotlivé receptory.



Obrázek 2.9: KIR geny a jejich vazebné ligandy. Pokud je v obrázku ? značí to, že pro daný receptor není znám vazebný ligand. [20]

2.5 KIR haplotyp

KIR haplotyp je vyjádření jaké konkrétní KIR geny genom obsahuje. Doposud nebylo zavedeno konkrétní pravidlo. Bylo navrženo, aby každý KIR haplotyp byl označen "*KH* – " následovaným trojmístným číslem, které bude označovat konkrétní haplotyp. Bylo by tak možné pojmenovat 999 haplotypů.

Dále by se haplotypy rozdělovali na dvě skupiny A a B. Skupina B musí obsahovat alespoň jeden z genů KIR2DL5, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5 a KIR3DS1. Naopak skupina A neobsahuje ani jeden z těchto genů. Z tohoto pravidla je patrné, že haplotypy B mají vždy více aktivačních KIR než haplotypy A. Za trojmístným číslem by tedy dále bylo písmeno A nebo B.

Nakonec by byl připojen 17-ti místný binární kód, který by označoval přítomnost "1" či absenci "0" genu. Pořadí genů by odpovídalo pořadí v genomu.












































































































Výsledné pojmenování by pak mohlo vypadat následovně:

$$KH - 001A - 11100010011011011 \quad (2.1)$$

TODO ale když se kouknu na ten obrázek co jsem stáhla tak to asi není úplně pravda protože mám i typy AB atd. Nesouhlasí mi to s tabulkou co mám dole, kterou jsem získala tady <https://www.ebi.ac.uk/ipd/>

kir/sequenced_haplotypes.html a ani přímo popis na <https://www.ebi.ac.uk/ipd/kir/haplotypes.html> k tomu nesedí. Kde jsem teda vzala AB?

TODO co je to za haplotyp který jste mi zaslala? Já ho v tý tabulce nemůžu najít.

Source	Haplotype	3DL3	2DS2	2DL2/3	2DL5B	2DS3/5	2DP1	2DL1	3DP1	2DL4	3DL1/S1	2DL5A	2DS3/5	2DS1	2DS4	3DL2
ABC08	A		—		—	—						—	—	—		
G085	A		—		—	—						—	—	—		
RSH	A		—		—	—						—	—	—		
FH06	A		—		—	—						—	—	—		
UV5HL9-5B	A		—		—	—						—	—	—		
LUCE	A		—		—	—						—	—	—		
FH13	A		—		—	—						—	—	—		
G248	A		—		—	—						—	—	—		
T7526	A		—		—	—						—	—	—		
FH05	A		—		—	—						—	—	—		
FH15	A		—		—	—						—	—	—		
FH08	A		—		—	—						—	—	—		
ABC08	AB		—		—	—									—	
GRC212	AB		—		—	—									—	
G085	BA1				—	—	—	—				—	—	—		
FH06	BA1				—	—	—	—				—	—	—		
GRC212	BA1				—	—	—	—				—	—	—		
FH08	BA2X		—									—	—	—		
FH13	BA2											—	—	—		
RSH	BA2											—	—	—		
G248	BA2											—	—	—		
LUCE	Bdel				—	—	—	—							—	
RPC1-1					—	—	—	—							—	
T7526	Bdel				—	—	—	—							—	
FH05	B														—	
FH15	B														—	
WT47	B														—	
Source	Haplotype	3DL3	2DS2	2DL2/3	2DL5B	2DS3/5	2DP1	2DL1	3DP1	2DL4	3DL1/S1	2DL5A	2DS3/5	2DS1	2DS4	3DL2
Key			Framework gene (can be activating, inhibiting or a pseudogene)					Inhibitory gene		Activating gene		Pseudogene	—	Gene not present		

Obrázek 2.10: KIR haplotypy. [17]

2.5.1 Dědičnost KIR

HLA je na 6. chromozomu KIR je 19 chromozomu. tudíž se segregují nezávisle a HLA shodní dárce s příjemcem mají obvykle různé složení KIR genů. Nesmírná variabilita alel tohoto systému ztěžuje úspěšnost allogeních transplantací.

Jelikož jsou geny kódovány na různých chromozomech (HLA 6 a KIR 19) takže HLA shodní dárce s příjemcem mají různé složení KIR genů. KIR má dva haplotypy A a B .. a pak asi můžeš dělat kombinace AA, AB a BB.

Tak jako geny HLA systému se i KIR geny dědí podobně a to jako celý blok genů – haplotyp (viz obr. 3).

A je tam k tomu hezkej obrázek disertačka Jindra

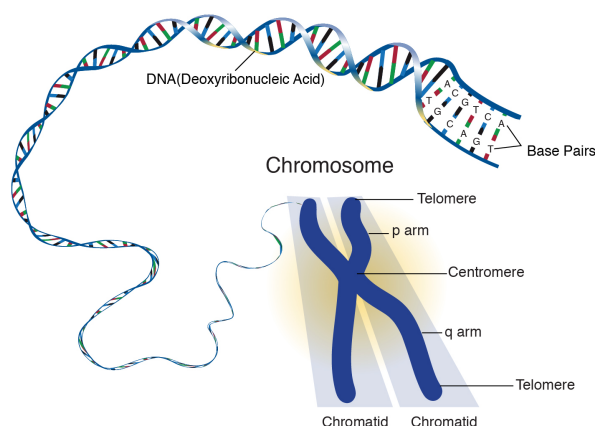
Bylo zjištěno, že specifické složení motivů centromerních a telomerních B haplotypů KIR genů přispívá k ochraně před relapsem a zvyšuje šanci na úplné vyléčení AML.

2.6 Sekvence DNA

Po pojmem sekvence DNA se skrývá posloupnost písmen představujících primární strukturu reálné nebo hypotetické molekuly či vlákna DNA, které nese nějakou informaci. Jednotlivá písmena jsou označována jako nukleotidy nebo nukleové báze. Nukleové báze mohou být A - adenin, C - cytosin, G - guanin a T thymin. [2]

Příkladem může být následující úsek sekvence na základě obrázku 2.11

$$ACGTCA \quad (2.2)$$



Obrázek 2.11: Umístění nukleotidů v lidském genomu. [1]

2.7 Alela a gen

Alelu můžeme definovat jako variantu genu s nepatrným rozdílem v sekvenci nukleotidů DNA oproti jiné alele stejného genu. Geny se vyskytují minimálně ve dvou formách (dvou alelách), mnohdy jich, ale může být více. U jednoho člověka mohou být přítomny pouze dvě rozdílné alely daného genu.

TODO: Když najdu novou sekvenci tak kde je rozdíl jestli je to nový gen nebo nová alela? Není to tak že na daný pozici v genu je vždycky gen.. a alela určuje tu vlastnost? A na co je mi teda lotus?

TODO tohle je asi jen HLA nevím jestli existuje něco jako obecné rozdělení genu a aleli, možná že rozdíl bude jen v tom že těch čísel pak může být za hvězdičkou několik v závislosti o alelu jaké skupiny genů se jedná Aleli jdou definovány HLA-DRB1* což označuje označuje lokus, následované 4 čísly. První dvojčíslí označuje sérologickou specifitu dané alely, druhé dvojčíslí označuje alelu na základě její aminokyselinové sekvence. Je možná se setkat ještě s pátým číslem které označuje záměnu nukleotidů bez změny aminokyselinové sekvence.

Teda že gen nám řekne: ano tenhle živočich má oči a alela mi řekne že oči jsou modré? možná do přílohy by se dalo dát jak vypadají dvě alely jednoho genu - to je asi blbost

Alela zajišťuje konkrétní fenotypový projev genu. U jedince mohou na homologních jaderných chromozomech být přítomny pouze dvě alely. Když jsou v párových lokusech obě alely shodné, jde buď o dominantního homozygota (AA) nebo o recesivního homozygota (aa). Když jsou na párových chromozomech v daném lokusu přítomny různé alely, jde o heterozygota (Aa). Značení alel vzniká dohodou. GEN je KIR3DP1 alela je *001 nebo *002

Zápis genů jak snimi budeme pracovat může vypadat následujícím způsobem: GTTCGGGAGGTTGGATCTGAGACGTGTTGTGAGTTGGTCA-TAGTGAAGGACGTGAGGTGC

KIR:KIR00087 KIR3DP1*001 5713 bp GTTCGGGAGGTTGGATCTGAGACGTGTTGTGAGTTGGTCATAGTGAAGGACGTGAGGTGC

2.7.1 Alely KIR genů

Když budu mít třeba gen KIR2DL1 tak jeho alely jsou 2DL1*0010101 nebo 2DL1*0010102

jestli to chápu tak gen mi určuje že budu mít oko a alela mi určuje jakou barvu to oko bude mít.

2.8 Bordel - KIR

haplotypická variabilita Tzv. „framework“ KIR geny (2DL4, 3DL2 A 3DL3, viz obr. 4) jsou přítomny prakticky u všech dosud typizovaných jedinců, ale přítomnost ostatních 14 KIR genů je výrazně variabilní. Tato variabilita v zastoupení je větší pro aktivační KIR geny („S“), které mají navíc limitovaný alelický polymorfismus. Inhibiční geny jsou obvykle vždy v genotypu přítomné, ale mají zase extenzivní alelický polymorfismus

Bylo popsáno celkem 15 exprimovaných KIR genů a 2 KIR pseudogeny

Pokud jde o vazebné partnery KIR (jejich ligandy), pak tyto jsou známy především pro inhibiční receptory a ve všech případech jde o HLA specifický I. třídy. Jedná se především o skupinu alel HLA-C alel lišících se aminokyselinným reziduem na pozicích 77 a 80 α - helixu molekuly HLA-C (7). Byla publikována rozsáhlá data ukazující na význam inhibičních KIR a jejich HLA ligand pro výsledek transplantace krvetvorných buněk

KIR geny/receptory a jejich vazební partneři (ligandy) - fryčová ta tabulka tam byla zajímavá, stejná je i v té disertačce od Jindry odtamtu to asi bude lepší

KIR geny jsou lokalizovány na chromosomu 19q13.4 v oblasti zvané „leukocyte receptor complex“ (LRC). Pro každý KIR gen navíc existují alelické varianty (Marsh et al, 2002;Hsu et al, 2002).

Genetická diverzita KIR genů a genotypů připomíná diverzitu HLA systému. Přestože jsou geny kódující KIR a HLA lokalizované na různých chromozomech a segregují se tedy nezávisle, existují určité důkazy alespoň částečné koevoluce obou systémů. Lze tudíž předpokládat. U HLA restrihované populace lze tedy očekávat alespoň částečnou redukci v diverzitě KIR genů i genotypů. U HLA restrihované populace lze tedy očekávat alespoň částečnou redukci v diverzitě KIR genů i genotypů.

Koevoluce je společný evoluční vývoj dvou či více druhů, při němž dochází k jejich vzájemnému přizpůsobování

Haplotypická variability KIR genů KIR geny se vyskytují ve dvou hlavních haplotypech A a B, které jsou definovány typem a počtem specifických KIR genů. Ta je způsobena variabilitou v počtu a v typu zastoupených KIR genů na daném haplotypu. Právě tato haplotypická diverzita je hlavním důvodem populační diverzity KIR genů a repertoáru NK buněk. Není žádné univerzální kritérium které by je odlišovali

Skupina B je charakterizována přítomností alespoň jednoho nebo více z následujících genů KIR2DL5, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5 a KIR3DS1.

Skupina A je charakterizována absencí těchto genů.

Proto mají B více aktivačních KIR než A. A může mít jen KIR2DS4.

2.9 bordel

Je posloupnost písmen představující primární strukturu reálné nebo hypotetické molekuly či vlákna DNA, které má kapacitu nést informaci. označuje se buď nukleotidy nebo nukleové báze. Používaná písmena A, C, G a T reprezentují čtyři nukleotidy ve vláknu DNA – adenin, cytosin, guanin a thymin, lišící se typem báze kovalentně vázané k fosfátové páteři. Posloupnost libovolného množství nukleotidů většího než čtyři lze nazývat sekvencí. Obvykle se sekvence vypisuje bez mezer, např. AAAGTCTGAC, ve směru 5' -> 3'. Vzhledem k biologickým funkcím, které mohou záviset na kontextu, sekvence buďto mají anebo nemají smysl a jsou tedy kódující nebo nekódující DNA. Typem nekódující sekvence DNA je také tzv. „junk DNA“. TO je z wiki bacha na to.

2.9.1 bordel hla

K typizaci se nejčastěji používá PCR – SSP (PCR se sekvenačně specifickými primery) či SBT (sequence based typing) technika, v posledních letech se stává zlatým standardem přímá sekvenace (SBT) HLA genu. fričová

Aby se určitý gen mohl fenotypově projevit (tj. přepsat se do určitého produktu), musejí se v jeho lokusu nacházet nejméně dvě alely – jedna pochází od matky, druhá od otce. HLA-antigeny však mají ve svých genových lokusech abnormálně velké množství alel, proto tvoří nejpolymorfnější systém v lidském těle.

antigen Jedná se v základu o rozpoznávací molekuly - jsou to vlastně proteiny kódované konkrétním genem, ve výsledku je to to stejné :), protein je vlastně vyjádření genu

HLA antigeny jsou přítomny na povrchu jsou přítomny na povrchu všech jaderných buněk

T-lymfocyty rozpoznávají cizí antigeny v komplexu s vlastními molekulami MHC, což vede k imunitní reakci, vlastní antigeny v komplexu s vlastními molekulami MHC, což vede k toleranci, cizí molekuly MHC (transplantační reakce).

Antigeny jsou velmi polymorfní. To znamená, že v lokusu určitého genu na daném místě DNA, kde se uvedený gen nachází, je možné najít minimálně dvě různé varianty tohoto genu. Ty se nazývají alely. K tomu, aby měl gen možnost se genotypově projevit, musí jeho lokus obsahovat dvě alely – jednu od matky a jednu otcovu. MHC lokusy obecně jsou geneticky nejvariabilnější kódující lokusy u savců a totéž lze tvrdit i o lidských HLA lokusech.

HLA antigeny je skupina bílkovin, jsou lokalizovány na povrchu buněk klíčové pro imunitní systém. Zjednodušeně lze říci že HLA molekuly, imunoglobuliny a receptory T-lymfocytů jsou základem adaptivní imunity člověka

HLA je rozsáhlý komplex genů, které determinují (určují, rozpoznávají???) povrchové molekuly (antigeny) umístěné v plazmatické membráně buněk

Hlavní fyziologickou funkcí molekul MHC je předkládat antigeny nebo jejich fragmenty buňkám imunitního systému, především T-lymfocytům (prezentace antigenu je prvním předpokladem pro rozvoj imunitní reakce a tím obrany proti napadení mikroorganismy). Pomocí těchto molekul buňky imunitního systému vzájemně kooperují.

2.9.2 bordel zbytek

exprimovaný gen / pseudogeny takže exprimovaný je ten který se propíše na povrch buňky a pseudogen je ten který se nepropíše na povrch buňky Pseudogen je sekvence DNA, která je podobná genu, ale nedochází k jejímu přepisování v RNA (transkripci). Pseudogeny vznikají zpravidla určitou mutací, a to zejména v oblasti promotoru, regulační sekvence,[1] nebo sice RNA vzniká, ale je následně degradována.

NK buňky mají schopnost identifikovat molekuly vlastního MHC sys-

tému (Major Histocompatibility Complex), jmenovitě HLA I. třídy, které jsou normálně exprimovány prakticky na všech buňkách v těle. Nádorové a některé virem napadené buňky potlačují expresi HLA I. třídy a tím se brání napadení cytotoxickými T lymfocyty (Restifo, 1993). Snížená exprese HLA I. třídy činí abnormální buňky citlivé k cytotoxicitě NK buněk (Karre, 1986). Molekuly HLA I. třídy rozpoznávají NK buňky pomocí pozitivních a negativních 13receptorů, které mohou inhibovat nebo naopak aktivovat NK buňky k „zabíjení“

Pokud to můžeme principálně zjednodušit, pak NK buňky neustále systematicky „zkoumají“ přítomnost či absenci příslušných HLA ligand pro své KIR receptory. Pokud je příslušný HLA ligand (HLA molekula) přítomen, pak dojde k vazbě KIR-ligand HLA a jelikož za normálních okolností vždy inhibiční KIR převládají nad aktivačními, nedochází ke spuštění cytotoxické reakce NK buněk a takto jsou „vlastní“ buňky chráněny před cytotoxicitou (viz část A a především D na obr. 1). Pokud receptory KIR nenaleznou příslušný ligand HLA („vlastní“ molekulu HLA), nemůže být cytotoxicita příslušné NK buňky prostřednictvím inhibičních receptorů KIR a náležitá cytotoxická kaskáda je spuštěna

Molekuly HLA I. třídy rozpoznávají NK buňky pomocí pozitivních a negativních receptorů, které mohou inhibovat nebo naopak aktivovat NK buňky k „zabíjení“

Tato schopnost destruovat cílové buňky je právě dána vzájemnou interakcí mezi KIR receptory a příslušnou specifickou HLA molekulou na povrchu buněk, neboli ligandem KIR receptorů

NK buňky neustále systematicky testují přítomnost či absenci příslušných HLA ligand pro své KIR receptory. Pokud je přítomen dojde k vazbě KIR-ligand HLA a nedojde k cytotoxické reakci (schopnost ničít buňky) , ochrana vlastních buněk. Pokud je nenajdou a je spuštěna cytotoxická kaskáda. Některé virem napadené buňky nebo nádorové buňky potlačují expresi a tím se brání napadení cytotoxickými T lymfocyty snížená exprese HLA I. třídy činí abnormální buňky citlivé k cytotoxicitě NK buněk.

Molekuly HLA I. třídy rozpoznávají NK buňky pomocí pozitivních a negativních 13receptorů, které mohou inhibovat nebo naopak aktivovat NK buňky k „zabíjení“

NK buňky mají schopnost identifikovat buňky vlastního MHC systému (HLA I.třídy) které jsou normálně exprimovány prakticky na všech buňkách v těle.

Lze-li stručně shrnout, pak NK buňky s potenciálem iniciovat cytotoxickou aloreakci používají KIRy jako inhibiční směrem k „vlastním“, zdravým

buňkám. Pokud však příslušný vlastní ligand HLA na cílové buňce chybí, pak je iniciována cytotoxická reakce. Celý tento koncept interakce KIR/HLA a mechanismus regulace cytotoxicity NK buněk se nazývá „missing-self“ hypotéza (3). Tím je zaručena tolerance NK buněk k „vlastním“ a zdravým buňkám, naopak alogenní („cizí“) buňky, či buňky s „down“- regulovanou HLA molekulou (ligandem), což jsou typicky buňky nádorové či buňky napadené virem, jsou efektivně eliminovány

Stručně lze shrnout, že NK buňky s potenciálem iniciovat cytotoxickou aloreakci používají KIR receptory jako inhibiční, směrem k „vlastním“, zdravým buňkám. Pokud však příslušný vlastní ligand HLA na cílové buňce chybí, pak dochází k iniciaci cytotoxické reakce. Proces interakce KIR/HLA a mechanismus regulace cytotoxicity NK buněk se jako celek nazývá 14. „missing-self“ hypotéza (Gasser a Raulet, 2006). Takto je zaručena tolerance NK buněk k „vlastním“ a zdravým buňkám, naopak „cizí“ alogenní buňky, buňky s „down“ – regulovaným HLA ligandem (molekulou), což jsou typicky buňky napadené virem a nádorové buňky, které jsou efektivně eliminovány

Tato schopnost destruovat cílové buňky je dána vzájemnou interakcí mezi KIR receptory a příslušnou specifickou HLA molekulou na povrchu buněk, neboli ligandem KIR receptorů.

2.10 Bordel pro první kapitolu

lymfocyty bílá krvinka je leukocyt - typ bílé krvinky - T a B lymfocyty - specifická imunita - NK buňky nespecifická imunita - vznikají v z lymfatických kmenových buněk v kostní dřeni Aha takže lymfatické řečiště je více propustné proto to co nejde do cév jde sem pak se to odfiltruje a pak se to vrací do krevního řečiště.

NK je v podstatě lymfocyt a to je typ bílé krvinky.

neutrofilní granulocyty jsou schopny vycestovat z kapilár do místa zánětu přeměněné monocyty přítomné v játrech v tělních dutinách (hrudní, břišní), ve slezině vy lymfatických uzlinách a kostní dřeni

KIR KIR Je to součástí genu - řadí se do přirozené (nespecifické) imunity narodil od B-buněk a T-buněk. - NK buňky představují 10-15% lymfocitů v periferní krvi - jsou to buňky které reagují rychle a efektivně likvidují především nádorové buňky a buňky infokované virem

NK nemají antigenné specifické receptory, jak rozeznávají abnormální buňky? NK buňky identifikují molekuly vlastního MHC systému

jmenovitě HLA I. třídy, které jsou normálně exprimovány prakticky na všech buňkách v těle. Nádorové a některé virem napadené buňky potlačují

expresi HLA I. třídy a tím se brání napadení cytotoxickými T lymfocyty (Restifo, 1993). Snížená exprese HLA I. třídy činí abnormální buňky citlivé k cytotoxicitě NK buněk (Karre, 1986). Molekuly HLA I. třídy rozpoznávají NK buňky pomocí pozitivních a negativních receptorů, které mohou inhibovat nebo naopak aktivovat NK buňky k „zabíjení“

Stručně lze shrnout, že NK buňky s potenciálem iniciovat cytotoxickou aloreakci používají KIR receptory jako inhibiční, směrem k „vlastním“, zdravým buňkám. Pokud však příslušný vlastní ligand HLA na cílové buňce chybí, pak dochází k iniciaci cytotoxické reakce. Proces interakce KIR/HLA a mechanismus regulace cytotoxicity NK buněk se jako

receptory imunoglobulinové (protilátka - protein, který je schopen jako součást imunitního systému identifikovat a zneškodnit cizí objekty - bakterie a viry) v těle. Protilátky jsou nositeli humorální imunity. Jsou to krevní bílkoviny vznikající v mízní tkáni. povahy nacházejících se na povrchu Natural killers buněk a některých T-buněk (Variabilita v sekvenci).

Struktura nukleových kyselin

jen skopírované z Nukleové kyseliny (polynukleotidy) jsou tvořeny dlouhými řetězci (mono)nukleotidů, vzájemně spojených fosfodiesterovými vazbami. Řadíme je k tzv.heteropolymérům, neboť jsou sestaveny z různých typů základních jednotek. Tato skutečnost je podstatná pro uchovávání a předávání informace, což je základní funkce nukleových kyselin v organismu. Homopolyméry (např. glykogen) obsahují pouze jeden typ monoméru (v našem případě glukózu), a tak nemohou plnit informační funkci.

3 Sekvenační metody získávání DNA dat

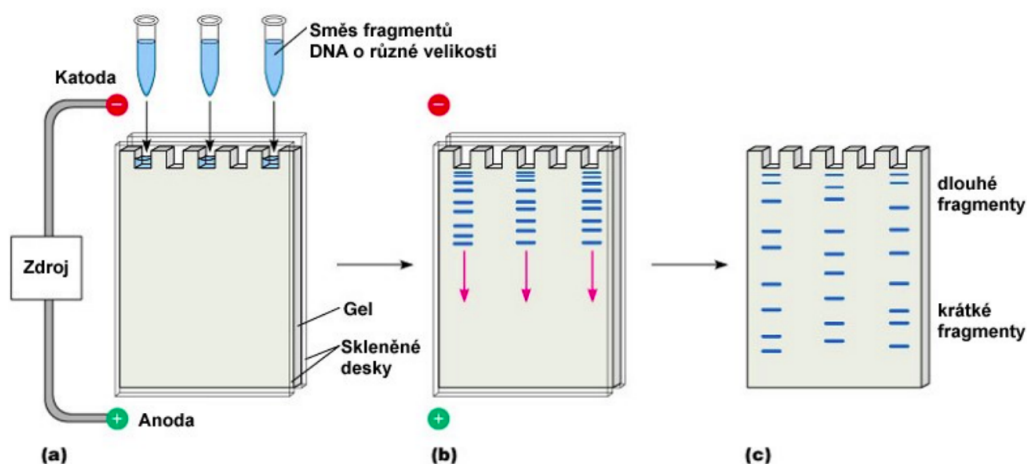
Sekvenování DNA, někdy pouze sekvenování, jsou biochemické metody, kterými se zjišťuje pořadí nukleotidů (A, C, G, T) v sekvenci DNA. Sekvenační metody se liší zejména délkou řetězce, kterou dokáží zpracovat, cenou a rychlostí sekvenace. Většina sekvenačních metod využívá vlastnosti přitahováním báze do páru pouze jednou konkrétní bází. To znamená že se adenin vždy páruje s thyminem a cytosin se vždy páruje s guaninem. Z těchto párů vzniká již známá dvojité šroubovice DNA. [10]

Dodat roztrhání a že délka řetězce taky hraje roly
možná tady ještě napsat něco o přípravě na sekvenování - je to dyžtak v
tý přednášce co nám říkala na FAV

Někdy se sekvenují pouze jisté části genomu které mají pro výzkumníka v daném okamžiku význam. Užitečné nejen ve výzkumu ale i v diagnostice nemocí či forenzní medicíně.

3.1 Sanger sequencing

Vzájemné přitahování konkrétních bází umožňuje řetězce namnožit. V prvním kroce replikace jsou nastříhané řetězce rozděleny na dvě vlákna. Lze si představit, že tyto dvě oddělená vlákna jsou dána do směsy, kde plavou jednotlivé nukleotydy spolu s upravenými nukletidy, které nesou specifickou fluorescenční barvu a na které a za které není možné nic navázat. Následně za pomoci střídání tepoty volně plující nukleotidy tvoří postupné páry s řetězcem, který chceme namnožit. Pokud se povede celý řetězec namnožit je odtržen a může se dále množit. Postupně ale bude docházet k navazováním nukleotidů s fluorescenční barvou. Tím se vytvoří několik různě dlouhých sekvencí zakončených označeným nukleotidem. Podle jeho barvy je možné poznat o jaký nukleotid se jedná. Následně jsou za pomoci elektorforézy seřazeny v gelu podle délky. Elektroforéza rozděluje různě dlouhé sekvence na základně odlišnosti pohybu v elektrickém poly. Kratší doputují dále než delší.



Obrázek 3.1: Elektroforéza. [14]

Mezi hlavní metodu patří sanger sekvenování oproti jiným je pomalé a spíše se používá k porovnání s novějšími metodami. TODO možná sem ještě přidat obrázek tý destičky.. ale zatím se nenašla nějaká hezká.

3.2 NGS next-generation sekvenování

Next-generation sekvenování někdy označováno jako metody druhé generace jsou v porovnání se Sangerovo sekvenováním rychlejší a levnější. NGS metody jsou schopné detakovat přidávání bází jednu po druhé a zároveň

jsou schopny detekovat přidávání bází jednu po druhé a zároveň sekvenovat tisíce až miliony rozdílných molekul DNA na jednou.

hlavní nevýhodou oproti sanger je krátká maximální délka výsledných sekvencí. od 100 až po 500 bází (sanger nabízí až 1000 bází) menší přesnost a častější chyby nejdříve nastříhané na malé krátké části na konec přilepen adaptér - velmi krátká molekula DNA o přesně dané sekvenci - slouží k následnému navázání sekvenovaného úseku na pevných povrhu. Takto upravené DNA se říká sekvenační knihovna po uchycení pomocí adaptéru je každý řetězec DNA namnožen čímž vznikne klastr identických molekul DNA koncentrovaných v jednom místě tato koncentrace posílí výsledný signál zachycený kamerou neboť signál z pouhé jedné molekuly DNA by nebyl dostatečně silný

Je rychlé a relativně nenáročné zpracování jednotlivých vzorků. Tisíce až miliony sekvencí mohou být produkovány během jednoho sekvenčního procesu. K popularitě této metody nepomohla i komerčializace cenově dostupných

stolních sekvenátorů.

3.2.1 454 sekvenování a Ion Torrent

454 bylo vypuštěno do světa v roce 2005. Dokáže analyzovat více než milion molekul DNA najednou a délka každé jednotlivé sekvence se pohybuje okolo 700 až 1000 bází.

V prvním kroku sekvenování je molekula DNA přichycena na malou "kuličku" na jejímž povrchu se postupně namnoží až kuličku zcela pokryjí identické molekuly DNA. Následuje vložení kuličky i s DNA do jedné z milionů komůrek na destičce s reakční směsí. V určitém momentě je do této směsi přidán vždy jen jeden typ báze. Mezi jednotlivými fázemi přidávání určité báze jsou přebytečné nukleotidy z předešlého kroku odstraněny. To znamená že v reakční směsi je vždy jen jeden typ nukleotidů. Během vložení každé nové báze do rostoucího řetězce DNA je uvolněna molekula zvaná pyrofosfát. Tento pyrofosfát se následně spustí několik chemických reakcí. V poslední fázi enzym luciferáza vydá světelný záblesk, který je možné zachytit citlivou kamerou. Tento postup se nazývá pyrosekvenování. V případě kdy je do řetězce přidáno několik stejných bází za sebou, například gen obsahuje podřetězec AAA, je vyzářeno, v našem případě, třikrát více světla než v případě jedné přiřazené báze. Kamera snímá celou destičku a na základě, která komůrka se rozsvítí pozná kde proběhlo přidání báze. Intenzita světla pak určuje kolik bází bylo přidáno na jednu.

Sekvenování Ion Torrent funguje na podobné principu sekvenování s rozdílem, že místo světla se měří změna pH v reakční směsi. Podle intenzity změny pH pak pozná kolik nukleotidů bylo přidáno do rostoucího řetězce.

Hlavní slabinou těchto dvou metod je značná chybovost při přidání mnoha stejných nukleotidů do řetězce za sebou. Například při přidání 10 A, nebude odpověď jednoznačná zda je to 10 A nebo 9.

3.2.2 Illumina

Při sekvenování pomocí Illumina jsou páry dvoušrobovice rozděleny na dva řetězce. Jednotlivé řetězce jsou následně přichyceny na malou destičku pomocí adaptéru. Každý řetězec se následně opakovaně množí až na destičce vznikne několik shluků. Přidání jedné molekuly ke druhé probíhá obdobně jako u Sanger sekvenování. Každý shluk tvoří jednu skupinu vzájemně iden-

tických řetězců. Mezi volné nukleotidy jsou opět zahrnuty nukleotidy označená fluorescenční barvou za které nelze nic navázat. Oproti sangerovu sekvenování je ale tato blokáce vratná a po přečtení citlivou kaměrou dojde k odstranění blukující části molekuly. Počítač si pak následně zpětně spočítá co to bylo za barvu (nukleotid.) Nejčastější chybou je špatně určené písmenko

Dokáže sekvenovat až 900 miliard? bází najednou. potřebuje kratší sekvence - stovky bází

Je tam hezká tabulka porovnání tak by se sem mohla taky dát

3.2.3 SOLiD

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) narozdíl od předchozích nespolehá na enzym DNA polymerázu, ale na enzym ligáza, který umí připojit části jednořetězcových molekul DNA k stávajícím řetězcům DNA.

Zjednodušeně lze říci, že při SOLiD sekvenování se k templátu přidají kousky DNA, tzv. sondy, které za - čínají všemi možnými dvojkombinacemi čtyř základních nukleotidů, tedy 16 různých sond. Každá sonda také nese jednu ze čtyř fluorescenčních značení, což znamená, že čtyři různé dvojkombinace nukleotidů jsou označeny stejnou fluorescenční značkou. V každém kroku pak enzym ligáza připojí k rostoucímu novému řetězci sondu nesoucí dvojkombinaci nukleotidů odpovídající templátové DNA a snímač přečte její fluorescenční značení, které je poté odstraněno a může se připojit další sonda. Aby došlo k přečtení kompletní sek - vence, je jedna templátová molekula čtena opakovaně, ale „začátek“ čtení se vždy posune o jeden nukleotid, a každá báze je tak přečtena několikrát. Z kombinace znalosti sekvence adaptéru, kterým sekvenovaná DNA začíná, a výsledného signálu čtyř fluorescenčních barev, jak jdou po sobě v jednotlivých čteních, lze odvodit výslednou DNA sekvenci. SOLiD sekvenování má podobný výstup jako Illumina a produkuje rovněž krátké sekvence (maximálně 100 bází). SOLiD má problémy se čtením palindromatických úseků (sekvencí shodných u obou komplementárních řetězců), jež mohou vytvářet smyčku v templátové DNA, která je pak nepřístupná pro nasednutí

3.3 Metody třetí generace

Narozdíl od drohé generace není DNA templat před sekvenováním nijak namnožen a tak je čten jen z jediné původní molekuly.

Například PacBio od Pacific Bioscience k detekci sekvence také využívá fluorescenčně značené nukleotidy. vysoká citlivost umožňuje v reálném čase zaznamenávat zařazení byť jediného nukleotidu do jediného řetězce DNA.

Oxford Nanopore Zde je jednořetězcová molekula DNA protahována mikroskopickým pórem na syntetické membráně. Protože každá DNA báze má trochu jiný tvar, dochází při protahování k odlišnému „ucpání“ póru a citlivé snímače přístroje dokážou zjistit, jak výrazně je pór v danou chvíli „zaplněn“, a tedy jaká báze v daný okamžik membránou prochází. výhoda je velikost .. je to malý kapesní přístroj který se dá přes USB připojit k počítači.

obě jsou schopné přechytit 10 i více tisíc bází v rámci jedné analyzované molekuly DNA a

vysoká frekvence chyb 10-15 procent

sekvenování celého genomu pomocí sangerova metody stálo několik miliard dolarů a trvalo zhruba 10 let. dnes by to stálo zhruba desítky tisíc dolarů

největší problém u sekvenování je že jsou roztrženy na malé části a pak je musíme zpět poskládat zpět.

zajímavost

Další metodou, která se dočkala rozmachu díky sekvenování druhé generace, je sekvenování transkriptomů (viz Živa 2016, 2: 61–63 a 3: 104–106). Při této metodě se místo kompletního genomu zjišťuje sek - vance pouze aktivních genů, tedy genů, které jsou v buňkách v danou chvíli přepisovány do mediátorové RNA (mRNA) a překládány do proteinů. Při sekvenování transkriptomů se nejdříve získá veškerá mRNA z daného organismu (nebo jen z určité tkáně, orgánu apod.) a přepíše se do DNA molekuly zvané copy DNA (cDNA). Teprve tato DNA je následně sekvenována. Výhoda sekvenace transkriptomů oproti celým genomům spočívá v získání sekvencí genů bez balastní (nepotřebné) nekódující části genomové DNA. Nekódující DNA totiž často tvoří podstatnou část genomu organismu a sekvence jednotlivých genů je proto potřeba v této záplavě pracně hledat, což se ne vždy spolehlivě daří.

Pomocí sekvenování transkriptomů také můžeme studovat odlišnou expresi (míru přepisu) jednotlivých genů v závislosti na vnějších nebo vnitřních podmínkách. Zpravidla porovnáváme transkriptomy získané z organismu nacházejícího se ve dvou či více různých „stavech“ – např. pěstování při různých teplotách nebo ve zdravém či nemocném stavu apod. Porovnáním přítomnosti a četnosti sekvencí jednotlivých genů lze určit, které geny jsou v příslušné „fázi“ aktivnější, a tedy nejspíše zodpovídají za reakci daného organismu na tento „stav“, třeba na změnu teploty nebo onemocnění.

Další novinkou vzešlou z dílny sekvenování druhé generace je sekve -

nování „exomu“. Místo celého genomu nebo v danou chvíli aktivních genů (transkriptomů) sek - venujeme pouze kódující část genomu, tedy geny jako takové, bez intronů (Pozn.: U většiny eukaryotických organismů se většina genů skládá z exonů a intronů. Exony představují kódující část genu, zatímco introny nic nekódují, a proto jsou před přeložením do patřičného genového produktu vystřiženy.). Samozřejmě předem potřebujeme velmi dobře znát genom daného organismu a hranice jednotlivých.

kódujících částí příslušných genů. Tato metoda se často používá při klinických studiích některých geneticky podmíněných chorob. V takovém případě lze porovnat exomy jedinců trpících poruchou a jedinců zdravých, a následně tak identifikovat mutace, které jsou pravděpodobně zodpovědné za nástup nemoci. Právě cenová dostupnost a vysoká efektivita metod sekvenování druhé generace dnes umožňuje zkoumat a rozpoznat příčiny řady vzácných a dříve málo studovaných genetických poruch.

3.4 Read

Read je sekvence bází odpovídající celému genomu či nějaké jeho části. Reads jsou typický výstup sekvenačních technik, kdy výstupem je sekvence nukleotidů o kterých nikdo neví co znamenají. Může to být gen, část genu nebo několik různých genů. Význam readu (o jaký gen se jedná) se zjišťuje zrovnáváním kdy se daná sekvence porovnává vůči referenčnímu genu.

In DNA sequencing, a read is an inferred sequence of base pairs (or base pair probabilities) corresponding to all or part of a single DNA fragment. A typical sequencing experiment involves fragmentation of the genome into millions of molecules, which are size-selected and ligated to adapters. The set of fragments is referred to as a sequencing library, which is sequenced to produce a set of reads. Je to z wiki zase

V DNA sekvenování, read je odvozená sekvece párů bází odpovídající celému fragmentu DNA nebo jeho části To znamená že read je kus DNA který by mohl odpovídat nějakého konkrétnímu genu?

Pak tam ještě bylo psaný něco o read lenght Sekvenační technologie se liší? v délce vyrobených readů. Reads díky 20-40 párů bází (bp) jsou ultrakrátké Typická sekvenační metoda vytváří reads délky 100 až 500 bp

DNA knihovny - podle wikiskripta DNA knihovny jsou kolekce klonovaných DNA fragmentů genomu určitého organismu (cDNA), které jsou sklado-

vány uvnitř hostitelských organismů (zejména bakterií). cDNA (copy DNA, complementary DNA) je získávána přepisem z mRNA pomocí enzymu reverzní transkriptázy.

Kvalita knihovny Při přípravě sekvenční knihovny je důležité získat co nejvyšší úroveň složitosti. Jinými slovy, je důležité, aby konečná knihovna co nejvíce odrážela jedinečnost výchozího materiálu. Tento výsledek lze získat především omezením počtu segmentových duplikací. Čím kratší jsou fragmenty, tím vyšší je pravděpodobnost, že jsou fragmenty méně specifické a mohou se zarovnat na více než jednom lokusu referenční sekvence. Složitost knihovny lze tedy v podstatě měřit procentem duplicitních čtení, které jsou přítomny v sekvenčních datech

Sekvenování mRNA s použitím NGS technologií umožňuje měření genové exprese celého transkriptomu. Postup a provedení RNA-seq experimentu je znázorněn na obr. 14. Prvním úkolem je vyčistit zkoumaný vzorek o rRNA, tRNA a mitochondriální RNA, které u prokaryot i eukaryot tvoří přibližně 75 procent všech RNA molekul. Navzdory použití purifikačních metod, mezi které patří například poly(A)purifikace a DNS normalizace, sekvenční data mohou obsahovat menší množství těchto RNA molekul [59]. Ty mohou být odfiltrovány v následujících krocích bioinformatickými postupy. Zbylá mRNA je poté nastříhána na menší části, a je z ní připravena knihovna krátkých fragmentů s navázanými adaptory. Ty jsou poté sekvenovány sekvenačním přístrojem a jako výsledek získáme tzv. ready. Anglické slovo 'read' značí datovou reprezentaci krátké sekvence DNA obvykle 50-150 bp dlouhou, která byla vyprodukována sekvenačním přístrojem. Samotné ready však nemají žádnou vypovídající hodnotu, a proto jsou dále bioinformaticky zpracovány. Namapováním na referenční sekvenci zjistíme jejich genomickou pozici, ze které byly odvozeny. Většina readů je namapována na exony, což jsou transkripčně aktivní jednotky, a pouze malé množství readů je namapováno na transposony. Ready které nejde namapovat v celku, jsou rozděleny na menší části a ty jsou namapovávány zvlášť. Rozdělené ready umožňují jednodušší identifikaci mezer mezi exony (angl. splice junctions) tohle je z té diplomky single-pair

3.5 Bordel

SANGER Vybraná sekvence se vloží do reakční směsi s radioaktivně označeným primer

Během procesu replikace jsou řetězce rozděleny na dvě vlákna. V praxi to není tak snadné a hrají roly i další proteiny. Mezi nukleotydy plavou i

upravené nukleotidy které nesou specifickou fluorescenční barvu a za ně už není možné aby s něčím nevázalo. Podle barvy poznáme o jakou bázi se jedná. Náhodným přerušováním syntézy vznikají různě dlouhé molekuly.

Výhody dlouhá délka sekvencí které se dají sekvenovat jedinou reakcí a vysoká přenosnost čtení v rámci celého procesu dochází k sekvenování pouze jednoho úseku DNA vysoká cena a nízká rychlost K sekvenaci se používá gelová elektroforéza použitelná k sekvenování krátké sekvence jednovláknové DNA. využívá biologického procesu replikace DNA Vybraná sekvence se vloží do reakční směsi s radioaktivně označeným primerem rozdělí se to na přibližně namnoží se to .. pak se to hodí do něčeho co nakonci svítí tak ty se navážou na příslušný konec.. pak to pustíme do gelu .. nejkratší projedou nejdál nejdelší zůstanou co nejbliž a podle toho pak sestavuju jak ta sekvence vypadá Sekvenování DNA je souhrnný termín pro biochemické metody, jimiž se zjišťuje pořadí nukleových bází (A, C, G, T) v sekvencích DNA. Tyto sekvence jsou součástí dědičné informace v jádru. Adenin s thyminem a cytosin s guaninem.

Zjišťování primární struktury nukleových kyselin (sekvenování) Někdy se sekvenují pouze jisté části genomu které mají pro výzkumníka v daném okamžiku význam. Užitečné nejen ve výzkumu ale i v diagnostice nemocí či forenzní medicíně.

454 , kde probíhá sekvenační reakce na principu pyrosekvenování.

Pyrosekvenování se založeno na skutečnosti, že během vložení každé nové báze do rostoucího řetězce DNA se uvolní molekula zvaná pyrofosfát (proto pyrosekvenování). Uvolněný pyrofosfát se posléze stane součástí několika na sebe navazujících enzymatických reakcí, na jejichž konci čeká enzym luciferáza Ten vydá světelný záblesk, jenž lze zachytit vysoce citlivou kamerou

Při 454 sekvenování je v určitém momentě přidán do reakční směsi vždy pouze jeden typ báze a v okamžiku, kdy je tato báze vložena do rostoucího řetězce DNA, dojde přes uvolněný pyrofosfát a luciferázu ke světelnému záblesku. Pokud je do rostoucího řetězce DNA zařazeno několik stejných bází za sebou, např. když DNA molekula templátu obsahuje sekvenci AAA a je tedy přidáno třikrát T, vyzáří se třikrát více světla než v případě přiřazení jednoho T

Kamera snímá celou destičku a podle toho, která komůrka se rozsvítí, pozná, kde proběhlo přidání báze, a podle intenzity světla kolik bází bylo přidáno najednou

Nukleotidy jsou přidávány jeden po druhém a mezi jednotlivými dochází k odstranění přebytečných nukleotidů. takže v reakční směsi je vždy jen jeden typ nukleotidů.

Ion Torrent je na podobném principu jako pyrosekvenování ale neměří světlo ale změnu pH v reakční směsi. Pod podle intenzity změny pH

protože spolehnají na sílu signálu aby věděli kolik bází bylo přidáno ne-jednou mají obě metody problém se čtením delších řetězců obsahující práce jen jednu bázi například AAAAAA. nebude jednoznačná odpověď zda je to 9 A nebo 10.

Illumina Dokáže sekvenovat až 900 miliard? bází najednou. potřebuje kratší sekvence - stovky bází pomocí adaptéru přichyceny molekuly DNA na malou destičku

Každá molekula DNA se pak opakovaně namnoží, až na destičce vznikne mozaika milionů klastrů, přičemž každou skupinu tvoří vzájemně identické molekuly. Vlastní sekvenační proces pak využije - vá podobného mechanismu jako Sangerovo sekvenování, kdy jsou do rostoucího řetězce zařazeny báze s navázanou fluorescenční barvou (každé písmeno má speci - fickou barvu), které syntézu zastaví.

tato blokáce je oproti Sangerovi vratná popřečtení citlivou kamerou dojde k odstranění fluorescenčního značení i blokující části molekuly a může se pokračovat

kamera snímá celou destičku a podle rozdílné fluorescence pozná co bylo přidáno u každého z milionů skupin

počítač si to pak zpětně přechrupa krok po kroku.

nejčastější chybou je špatně přečtené písmenko. jinak má 99 procentní úspěšnost

4 Analyza dostupných bioinformatických nástrojů pro zpracování NGS dat

4.1 Referenční geny

Referenční geny byly převzaty z IPD-KIR [16] konkrétně soubory ve formátu *fasta* uloženy ve stejnojmenné složce. Jednotlivé soubory jsou pojmenovány genem který obsahují např. *KIR2DL1_gen.fasta*. Každý soubor představuje všechny dostupné alely konkrétního genu.

Kromě souborů **_gen.fasta* obsahuje složka *fasta* také soubory **_prot.fast* a **_nuc.fasta*. Soubor **_prot.fast* obsahuje sekvenci proteinů, *nuc* obsahuje nucleotidy a *gen* obsahuje genomickou DNA sekvenci

TODO já vlastně nechápu jaký je rozdíl mezi *nuc* a *gen*?

All files in this folder are provided in the FASTA sequence format. Please note the FASTA format contains no alignment information.

Files designated “X_prot.fasta”, where X is a locus or gene, contain protein sequences. Please note that alleles that contain non-coding variations may be identical at the protein level.

Files designated “X_nuc.fasta”, where X is a locus or gene, contain the nucleotide coding sequences (CDS). Please note that alleles that contain non-coding variations may be identical at the CDS level.

Files designated “X_gen.fasta”, where X is a locus or gene, contain genomic DNA sequences. Please note for alleles that do not possess genomic sequences, there will be no entry in the file.

4.1.1 Vytvoření testovacího haplotypu

TODO navíc hned když se podívám na 3DL3: 00402 tak tam mám dvě možnosti TODO a co když tam nějaký není, tak prostě pokračuju tím dalším, nebo se tam něco dává jako mezera? Myslím když jste třeba napsala 2DL5B: TODO Co je 2DP1, ve FASTA jsem je nenašla tak jsem to skopčila z https://www.ebi.ac.uk/cgi-bin/ipd/kir/get_allele.cgi?2DP1*0020103 - p

značí pseudogen A prej by měli být v KIR-gen.fasta ale já to tam za boha nemůžu najít jo tak tam je v tom KIR_gen.fasta ale nahoře jsou i jiný co asi úplně nejsou pseudogeny tak to moc nechápu ..

to děláte správně, jen musíte vytvořit KIR haplotyp, který podšoupnete tomu ARTu, nejenom jeden konkrétní KIR, aby Vám vytvořil ready, tj. např. kombinaci (tople je jedna známá linie, sloučíte si ty KIRy za sebe):
3DL3: 00402, 00802 2DS2: 00101 2DL2: 00301 2DL3: 001 2DL5B: 2DS3:
2DP1: 00201 2DL1: 00302 3DP1: 007, 00901 2DL4: 00102, 00501 3DL1:
01502 3DS1: 01301 2DL5A: 001 2DS3: 2DS5: 00201 2DS1: 00201 2DS4: 001
3DL2: 0020105, 0070102

4.2 ART

ART (next-generation sequencing read simulator) je sada simulačních nástrojů, které generují syntetické ready, jako kdyby byli získány sekvenováním pomocí NGS. Nástroj ART dokáže simulovat ready ze sekvenátorů Illumina, 454 společnosti Roche a SOLid od společnosti Applied Biosystems. Ready, vytvořené nástrojem ART jsou používány pro testování a analýzy nástrojů zpracovávající právě NGS sekvence jako například zarovnávání (nástroj Bowtie).

ART je implementován v jazyce C++ a je dostupný s licencí GPL verze 3 pro operační systémy Linux, MacOS a Windows. Je možné ho použít i jako C++ package. Pro jeho spuštění je nutné mít nainstalovaný kompilátor GNU g++ 4.0 nebo vyšší a knihovnu GNU gsl.

Data získána z FN Plzeň byla sekvenována nástrojem Illumina proto i syntetické ready budou simulovat tento sekvenátor. Výstupy se čtou ve formátu FASTQ a zarovnání ve formátu ALN. může generovat zarovnávání také ve formátu SAM nebo UCS BED. [8]

4.2.1 pokus to nějak spustit

Takže když otevřu hlavní readme tak mi to říká že tam jsou read me pro jednotlivé verze sekvenátoru ..

pak se to musí skompilovat

./configure --prefix=\$HOME make make install

teď mě zajímá ta Illumina tak podle readme Illumina tak můžu vlést do složky examples a tam pustit skript run_test_examples_Illumina.sh , tak

tam jsou 4 příklady použití a pokud asi všechno dobře proběhne tak se mi zobrazí pár nových souborů ve složce examples..

FASTQ - *.fq data file s ready. pro paired-end simulator *1.fq obsahuje data pro první ready a *2.fq rdu druhý ready

tohle nějak funguje MSv3 tam musím dát abych to mohla dostat na délku readu 250 a p znaci ze to je paired.. tak se má používat MSv1 *art_illumina-ssMSv3-sam-iamplicon_reference.fa-p-l250-f10-m300-s10-omoje_art_data* Tohle používej: *art_illumina-ssMSv1-sam-iamplicon_reference.fa-p-l250-f100-m300-s10-omoje_art_data*

4.2.2 FASTQ

Sekvenační přístroje produkují data ve formátu FASTQ takže i ART musí logicky generovat tenhle formát. Pokud jsou ready v páru tak je na konci .1 a druhý read z páru tam má .2 to jsem u těch svých přímo nenašla

ale máš teda tři druhy single end, paired-end a matepair.

FASTQ obsahuje obě základy sekvence ?? both sequence bases a kvality skore je to v následujícím formátu @read_id sequence read + base quality scores je kódovány by ascii code of a single character, kde je kvalita rovná score to ascii code character minus 33. chápu proč tam je to -33 protože když se podíváš do asci tabulky tak je tam od 33 první normální znak jinak jsou tam divný .. takže třeba otazník je v asci na 63 takže -33 takže má ohodnocení kvality 30 jen by mě teda zajímalo v jakém sme intervalu? - je 45 v asci a nevím jestli to je teda od 0 do 100? a teda nejvyšší číslo znamená nejkvalitnější a nejmenší mín kvalitní? Podle té diplomky to tak je že čím vyšší číslo tím kvalitnější a většinou je to od 0 do 40 jen zřídka to překročí hodnotu 60, když je tam 10 tak to znamená že jedna báze z deset je špatně.. když je tam 30 tak to znamená že jedna z 1000 je špatně. já tam mám třeba F a to je 70.

example: @refid-4028550-1 caacgccactcagcaatgatcggtttattcacgat... +

ALN - zarovnání readů zase *1.aln pro první a *2.aln pro druhý soubor je rozdělen na hlavičku a body část obsahuje hlavičku a v té hlavičce je jakým příkazem byl soubor vygenerován a reference na seunce id a jejich délku @CM tag pro příkaz a @SQ pro reference sequence Hlavička vždycky začíná s

HEADER EXAMPLE

v body jsou všechny zarovnání

aln_start_pos označuje počáteční pozici v referenci sekvence, je vždy relativní vzhledem k vláknům referenční sekvence To znamená že aln_start_pos plus (10) vláknem je odlišný od aln_start_pos minus (-) vláknem.. ??? WHAT???

ref_seq_aligned je zarovnaná oblast referenční sekvence, která může být plus vlákno nebo mínus vlákno referenční sekvence ref_seq_aligned je zarovnaný read, který je vždy ve stejné orientaci jako stejný read v odpovídajícím fastq suboru.

aln_start_pos is the alignment start position of reference sequence. aln_start_pos is always relative to the strand of reference sequence. That is, aln_start_pos 10 in the plus (+) strand is different from aln_start_pos 10 in the minus (-) strand.

ref_seq_aligned is the aligned region of reference sequence, which can be from plus strand or minus strand of the reference sequence. read_seq_aligned is the aligned sequence read, which always in the same orientation of the same read in the corresponding fastq file.

SAM je standardní formát pro NG sekvence ready zarování BED o tom tam nic není jen NOTE: both ALN and BED format files use 0-based coordinate system while SAM format uses 1-based coordinate system.

pak jsou tady 4 doporučené použití *art_illumina[options] -ss < sequencing_system > -sam -i < seq_ref_file > -l < read_length > -f < fold_coverage > -o < outfile_prefix > art_illumina[options] -ss < sequencing_system > -sam -i < seq_ref_file > -l < read_length > -c < num_reads_per_sequence > -o < outfile_prefix > art_illumina[options] -ss < sequencing_system > -sam -i < seq_ref_file > -l < read_length > -f < fold_coverage > -m < mean_frag_size > -s < std_frag_size > -o < outfile_prefix > art_illumina[options] -ss < sequencing_system > -sam -i < seq_ref_file > -l < read_length > -c < num_reads_per_sequence > -m < mean_frag_size > -s < std_frag_size > -o < outfile_prefix >*

pak tam máš parametry

a jak dlouhý chceme simulovat ready?

4.2.3 bordel

ART is freely available to public. The binary packages of ART are available for three major operating systems: Linux, Macintosh, and Windows. ART is also available as Platform-independent C++ source packages. Each package includes programs, documents and usage examples.

ART simuluje ready napodobobáním skutečných procesů sekvenování s empirickým chybovým modelem nebo quality profiles summarized from large recalibrated sequencing data ART může také simlovat čtené pomocí uživatelského vlastního read error modelu nebo quality profiles

TODO - tohle úplně nechápu ART podporuje simulaci jedno párových, dvou párových tří hlavních komerčních sekvenčních platfoem Výstupy se

čtou ve formátu FASQ a zarování ve formátu ALN. ART může také generovat zarovnávání ve formátu SAM nebo UCSC BED ART lze použít společně se simulátory variant genomů VarSim

to je odtud 454 sekvenování je pyrosekvenování, které cyklicky testuje přítomnost každého ze čtyř nukleotidů DNA (T, A, C, G)

SOLid ke kódování 16 různých dinukleotidů používá čtyři fluorescenční barevná barviva, každé barvivo kóduje čtyři dinukleotidy

tak jsem stáhla normálně nejnovější verzi z niehs.nih.gov a podle instrukcí co byli v souboru INSTAL dala

musí se brát v potaz že z toho generátoru nikdy nebudou data taková jako reálná.. reálná budou horší

4.3 Bowtie

Bowtie je rychlý a paměťově efektivní nástroj pro zarovnávání krátkých sekvencí DNA na velké genomy. Indexace pomocí Burrows-Wheelere transformace dovoluje zarovnávání více než 25 milionů readů za CPU hodinu pro lidský genom s pamětí přibližně 1.3 gigabajtů. Bowtie přidává k Burrows-Wheeler technice backtracking algoritmus pro sledování nekonzistence. ??

4.3.1 Bordel

Bowtie je napsaný v c++ a používá knihovnu seqAn

Na lidském genomu je nástroj Bowtie v porovnání s nástroji Maq a SOAP rychlejší. Citlivost má bowtie srovnatelné s nástrojem SOAP a o něco menší než Maq. Ale je možnost pomocí příkazové řádky zvýšit citlivost na úkor rychlosti běhu programu. Oproti SOAP bowtie potřebuje méně paměti 1.3 GB RAM. Bowtie zarovnává 25 milionů readů za hodinu. může běžet paralelně.

indexi vytváří permanentní a lze je použít napříč běhy pro lidský genom je to 2.2 GB takže ho lze distribuovat přes internet rychlost a malá paměť způsobuje především Burrows wheeler v kombinaci s backtrackingem.

Podporuje standardní vstupní formáty FASQ a FASTA.

Bowtie je open source.

na stránkách [elixir-europe](http://elixir-europe.org) což je organizace co má dávat dohromady všechny vědecké věci a bla bla.

Tak tam je přímo Bowtie [12]

4.3.2 Bowtie 2

Note that SOAP2 and Bowtie do not permit gapped alignment of unpaired reads. memory footprint of Bowtie 2 (3.24 gigabytes) Bowtie 2 by mělo být vhodnější pro delší reads než Bowtie1. We extracted a random subset of 1 million reads from each and aligned them with BWA-SW and Bowtie 2. We did not align with Bowtie, BWA or SOAP2 because those tools are designed for shorter reads. Bowtie už je překonané nejenom Bowtie2 ale i BWA. Bowtie2 je podle studie znatelně lepší než Bowtie, SOAP2. tyto výsledky jsou na syntetických readech

vypadá to že bowtie 2 už nepoužívá tamten index ale používá nějaký Full-text minute index-assisted search což vypadá že je kombinace burrows wheelera a ještě něčeho. We found that Bowtie 2, a method that combines the advantages of the full-text minute index and SIMD dynamic programming, achieved very fast and memory-efficient gapped alignment of sequencing reads

je zase open source [11]

šla jsem přes docker docker image ls - zobrazí všechny image pak docker run a ID image sudo docker run -i -t 3c2b9a287f82 /bin/bash sudo docker ps -a

Tak jsem nakonec žádné docker nepotřebovala a stáhla jsem to tady po kliknutí na bowtie binary release.

na stránce 25.4 je řečeno o hledání těch nejlepších zarovnání a je tam možnost -best ale že je dvakrát nebo třikrát pomalejší než normální mod.. a jde o to že najde první přijatelný a to označí kdežto při tom best prohledá co nejvíce a hledá to nejlepší i mezi těmi přijatelnými a to je pomalý.

takže zarovnání by mohlo být teoreticky namapování na referenční gen???

4.3.3 bordel

tak jsem to stáhla dala do složky a musela jsem teda nastavit proměnou prostředí export BT2_HOME=/home/kate/Dokumenty/FAV/Diplomka/existujicisw/bowtie2-2.4.1-linux-x86_64/ pak jsem pustila tohle: \$BT2_HOME/bowtie2-build \$BT2_HOME/example/reference/lambda_virus.falambda_virus a nakonec se mi vytvořili nějaký nové soubory lambda virus 1 atd.. v tom bowtie 2 adresáři

dělala jsem to podle tohohle webu

z bowtie pak teda leze asi SAM formát

SAM

1. název readu který je zarovnáván

2. Sum of all applicable flags. Flags relevant to Bowtie are: součet všech aplikovaných (příslušných flags). Flagy relevantní k bowtie jsou: 1 - read je jeden z páru 2 - zarovnání je one z paired proper (The alignment is one end of a proper paired-end alignment) 4 - read má reported alignments 8 - read je jeden z páru a má reportovaný zarovnání 16 - zarování je obrácená reference vlákna 32 - The other mate in the paired-end alignment is aligned to the reverse reference strand 64 - read je mate 1 in a pair 128 - read je mate 2 in a pair

Thus, an unpaired read that aligns to the reverse reference strand will have flag 16. A paired-end read that aligns and is the first mate in the pair will have flag 83 ($= 64 + 16 + 2 + 1$).

3. jméno referencce ze které zarování patří 4. 1-based offset into the forward reference strand where leftmost character of the alignment occurs 1-based odszaneí v následující referenci 5. kvalita mapování 6. CIGAR reprezentace zarovnání 7. název reference kde je zarovnán kamarád 8. 1-based zarování ofsetu k následující referenci 9. Odvozená délka fragmentu. Velikost v závorku je že se mate nachází předtím. 0 že jsem nezarovnali mate 10. read sekvence 11. ASCII encoded read kvalita, stejné jako u FASTQ 12. optional pole

SAM Sequence Alignment Map format), respektive jeho binárně komprimovaná verze BAM (z angl. Binary Alignment Map format).

nakonec jsem to pustila přes IGV ale stejně se tam museli ty indexi dodělat

5 Seznam zkratek

HLA KIR NK sekvence DNA cytosin, adenonin atd DNA (Deoxyribonukleovou kyselinu) co ty formáty souboru?

6 Výkladový slovník pojmů?

WHO český národní registr možná zmínit národní registr genotyp fenotyp
tyhle kraviny

Literatura

- [1] *Chromosome* [online]. [cit. 2020/12/3]. Dostupné z: <https://www.genome.gov/genetics-glossary/Chromosome>.
- [2] *DNA sequencing Fact Sheet* [online]. [cit. 2019/03/1]. Dostupné z: <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet>.
- [3] *S transplantací kostní dřeně stále častěji pomáhají příbuzní* [online]. Dostupné z: <https://ct24.ceskatelevize.cz/domaci/2527141-s-transplantaci-kostni-drene-stale-casteji-pomahaji-pribuzni>.
- [4] *Basic genetics* [online]. [cit. 2020/12/3]. Dostupné z: <https://kintalk.org/genetics-101/>.
- [5] BERNAREGGI, D. – POUYANFARD, S. – KAUFMAN, D. S. Development of innate immune cells from human pluripotent stem cells. 2019. Dostupné z: <https://www.sciencedirect.com/science/article/pii/S0301472X19300037?via%3Dihub>.
- [6] FRYČOVÁ, M. Lze u pacientů s AML indikovaných k nepříbuzenské transplantaci provádět v klinické praxi výběr nepříbuzných dárců na základě KIR genotypů, 2016.
- [7] HERNYCHOVÁ, L. Receptory NK buněk. 2012.
- [8] HUANG, W. et al. ART: a next-generation sequencing read simulator. 2012. Dostupné z: <https://academic.oup.com/bioinformatics/article/28/4/593/213322>.
- [9] J, R. et al. *Nomenclature* [online]. Nucleic Acids Research, 2015. [cit. 2019/10/1]. 43:D423-431. Dostupné z: <http://hla.alleles.org/misc/citing.html>.
- [10] KOLÍSKO, M. Moderní metody sekvenování DNA. 2017. Dostupné z: <https://ziva.avcr.cz/files/ziva/pdf/moderni-metody-sekvenovani-dna.pdf>.
- [11] LANGMEAD, B. – SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. 2012. Dostupné z: <https://www.nature.com/articles/nmeth.1923>.
- [12] LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. 2009. Dostupné z:

<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25>.

- [13] MUDR. PAVEL JINDRA, P. D. *Imunopatologické a imunogenetické aspekty transplantací krvetvorných buněk a solidních orgánů*. PhD thesis, Universita Karlova v Praze, 2011.
- [14] PAPOUŠEK, I. Elektroforéza nukleových kyselin. 2017. Dostupné z: https://fvhe.vfu.cz/files/mbhp_2018_02.pdf.
- [15] PENKA, M. – KOLEKTIV, E. T. *Hematologie a transfuzní lékařství II*. 2012. ISBN 978-80-247-3460-6.
- [16] ROBINSON, J. et al. The IMGT/HLA Database. 2013. Dostupné z: <https://www.ebi.ac.uk/ipd/index.html>.
- [17] ROBINSON, J. et al. IPD—the Immuno Polymorphism Database. 2013. Dostupné z: <https://www.ebi.ac.uk/ipd/index.html>.
- [18] S.KANNANA, G. – ARIANEXYSAQUINO-LOPEZ – A.LEED, D. Natural killer cells in malignant hematology: A primer for the non-immunologist. 2017. Dostupné z: <https://www.sciencedirect.com/science/article/pii/S0268960X16300704>.
- [19] SMITH, D. T. *Encyklopedie lidského těla*. 2005. ISBN 80-7321-156-4.
- [20] THIELENS, A. – VIVIER, E. – ROMAGNÉ, F. NK cell MHC class I specific receptors (KIR): from biology to clinical intervention. *Current opinion in immunology*. 2012, 24 2, s. 239–45.